11/22/2017

SUBMITTED TO: DR. READ

Submitted by: sukhdeep kaur

BIG DATA AND CLOUD COMPUTING

ASSIGNMENT 04

**BIG DATA**

Big data is mainly refers to a huge amount of data. Modern information technology is leading to development of all walks of life. Traditional database system is no longer able to solve various problems such as data collection, storage and analysis which is a main concern today. Big data has been defined in different ways one of which is given below

Big Data refers to datasets whose size is beyond the capability of typical database software tools to capture, store, manage, and analyze. —McKinsey.

Big data has four characteristics:

1. Volume: huge amount of data
2. Velocity: fast processing speed
3. Variety: various data types
4. Value: unlimited commercial value

**CLOUD COMPUTING**

Cloud computing is offering high availability, high scalability and high reliability. Cloud computing is a large scale distributed computing which provides storage networking to user in service mode according to their need. Most of the companies today are adopting cloud computing as the public cloud services from Amazon, Google and Microsoft become better developed.

Several technologies are considered the root of Cloud Computing.

1. Grid and Utility Computing
2. Virtualization
3. SOA and Web Services
4. Autonomous Computing

**DIALECTICAL RELATION BETWEEN BIG DATA AND CLOUD COMPUTING**

Cloud computing and big data are compatible, forming a relation-ship. The former is a dream of humanity’s pursuit of civilization, the latter is the bottleneck to be solved in social development. Big data is a phenomenon of rapid development in modern IT, while cloud computing is an increase in technology development. To solve big data problems, we need cloud computing technologies.

From the development of IT, few patterns can be summarize:

1.Modern IT technologies need to adopt the concept of openness.

2.Cloud computing is based on distributed architecture of open platform which allow it to solve problems that are difficult to solve with existing centralised approach.

3.The traditional industry giant is migrating towards open source architecture to achieve breakthrough innovations, defeat monopolies, and catch up with international advanced technologies.

**BIG DATA TECNOLOGIES:**

* Infrastructure support : Platform, cloud storage and virtualisation technology
* Data acquisition : Data bus and ETL Tools
* Data storage : Distributed File System, Relational database, NoSQL
* Data computing : Data mining and prediction, Graph analysis
* Display and interaction : Graphics and reports, visualisation tools

**1. INFRASTRUCTURE SUPPORT FOR BIG DATA 🡪 CLOUD COMPUTING PLATFORM**

Infrastructure support mainly includes infrastructure-level data centre management, platforms, storage equipment and technology, networking technology and resource monitoring technology. Big data processing needs a support from cloud data centres. Cloud computing management platforms can provide flexible and efficient deployment, operation and management environment for large data centres.

1.1 **Cloud computing platforms:** Amazon Web Services(AWS), Google’s app engine and Microsoft’ Windows Azure services are well known cloud servers. Each platform has its own significance and its own features. AWS is the most popular cloud computing platform: in the first half of 2013, its platform and cloud computing services have earned $1.7 billions with 60% annual growth. The most significant features of its system are open data and SOA(Service Oriented Architecture). The services provided by AWS can be divided into four layers:

* + 1. The Access Layer: It provides management console, API and various command-line tools.
    2. The Common Service Layer: This includes authentication, monitoring, deployment, and automation.
    3. The PaaS Layer: Includes parallel processing, content delivery, and massaging services.
    4. The IaaS Layer: Includes computing platform EC2, cloud storage services S3/EBS, network services VPC/ELB and database services.

1.2 **Cloud storage and virtualization:** Open Nebula is an open source implementation of the virtualisation management of virtual infrastructure. Open Nebula has integrated storage, network, virtualization, monitoring, and security technologies.

Another open source of virtualisation is OpenStack which allows users to run their storage infrastructure. OpenStack is among the best for the implementation of SOA and decoupling of service oriented components. OpenStack has mainly four layers:

1.2.1 Access Layer: Applications, management portals and APIs.

1.2.2 Core Layer: Computing services, storage services and network services

1.2.3 Shared services: Identity management services and image service(Glance).

**2. DATA ACQUISITION:** Data acquisition can be accomplished via sensors in the Internet of Things and can derived from network information. First we need means of data acquisition for collecting the information and then apply top layer data processing technologies. ETL is preprocessing which includes cleaning and filtering.

**3. DATA STORAGE:** Facing the huge amount of data distributed file system is used to store data in different nodes. Cloud storage, google file system and MapReduce distributed computing are basic tools to manage storage. Google’s technology was not open source, so Yahoo and Open source communication developed Hadoop system which is an open source implementation of GFS and MapReduce.

**4. DATA COMPUTING:** Data queries, statistics, analysis, mining, and other requirements for big data processing have motivated different computing models of big data, and we divide big data computing into three categories:

4.1 Offline batch computing: The Hadoop is mainly for offline batch applications and is typically used to batch tasks on static data.

4.2 Realtime interactive computing: The real-time computing process of massive data can be divided into the following three phases – real-time data collection, real-time data analysis and processing, real-time query services.

4.3 Streaming computing: Stream computing is designed for real time and continuous data, analyzing the movement process in real-time while the stream data is changing.

**5. DATA PRESENTATION AND INTERACTION:** Tableau, a big data start-up company from Stanford, is becoming one of the most outstanding data analysis tools. Tableau combines data computing and aesthetic charts perfectly.

**CONCLUSION:**

Big Data problems need to be solved by Cloud computing technology, while big data can also promote the practical use and implementation of Cloud computing technology. There is a complementary relationship between them. We focus on infrastructure support, data acquisition, data storage, data computing, data display, and interaction to describe several types of technology developed for big data.